A comparison of NIPALS algorithm with two other missing data treatment methods in a principal component analysis.

Abdelmounaim KERKRI* Zoubir ZARROUK** Jelloul ALLAL*

*Faculté des Sciences – Université Mohamed 1er, BV Mohammed VI BP 717 60000 Oujda

**Faculté des Sciences Juridiques, Economiques et Sociales - Université Mohamed 1er

Abstract

Missing data is a frequent problem in almost every survey; it can interfere with the quality of the statistical analysis in terms of bias, and therefore several methods of dealing with this problem have been developed. In this paper we present NIPALS algorithm, and two other existing methods of imputation, and we compare their performances in a principal component analysis using a simple indicator of square differences of the eigenvalues. The comparison is conducted on simulated data on which we vary the percentage of missing values from 5 % to 20% , the comparison shows the stability of this indicator when we use the bootstrap expectation maximization to impute the data instead of the NIPALS algorithm or the naïve mean imputation method.

Keyword:NIPALS, EM algorithm, missing data, PCA

Résumé

Le problème des données manquantes est fréquent dans presque tous les sondages ; il peut influencer négativement la qualité de l'analyse statistique, et donc plusieurs méthodes problème ont été développées ces dernières années. Dans cet article, nous comparons la performance de l'algorithme NIPALS par rapport a deux autres méthodes d'imputations (l'algorithme expectation maximisation et la méthode d'imputation par la moyenne) dans le cas d'une analyse en composantes principales selon un simple indicateur des différences des carrées des valeurs propres. La comparaison est effectuée sur des données simulées sur lesquels nous faisons varier le pourcentage de valeurs manquantes de 5% à 20%, la comparaison montre la stabilité de cet indicateur lorsque nous utilisons l'algorithme expectation maximization avec bootstrap pour imputer les données au lieu de l'algorithme de NIPALS ou la simple imputation par moyenne.

Mots clés : NIPALS, l'algorithme EM, données manquantes, ACP

Introduction

Several techniques to deal with missing data have been developed in the past few years [1,2,3]; some are considered to be naïve and outdated such as complete case analysis which only analyses the existing cases, simple imputation which consists of filling the missing cases with the mean or the median of the observed values of the same variable which is commonly used because of its simplicity. The single imputation uses a simple linear regression on the existing values, which can underestimate the variability of the data. Classification algorithms are also widely used such as k-nearest neighbor which imputes the mean or the median of the appropriate class; other methods are thought to be more robust in terms of performance, such as the expectation maximization algorithm [5], and the predictive models which consist of building a model to estimate the missing values. For categorical data, there are algorithms based on decision trees such as C4.5 [12], The performance of these different methods is usually tested by their output's similarity to the complete data set before the modeling process [11], the drawback in this case is that we don't test the effect of the imputed data set in the data mining process, that's why we chose to compare the three methods by their influence after the data mining process (principal component analysis).

In this paper we start by presenting the different types of missing data patterns. Then we present three treatments to compare; the main algorithm in this study is NIPALS, which performs a principal component analysis on incomplete data without having to estimate the missing values; this algorithm is the foundation of partial least squares regression models. Our choice was motivated by the fact that NIPALS is not very commonly used in the statistical community; the other two methods are the EBM algorithm [8], which is a slightly improved version of the classical expectation maximization algorithm implemented in the R software, and the simple mean imputation which, despite of its drawbacks, is commonly used to overcome the missing data problem.

1. Missing data

It occurs when no data value is stored for the variable in a certain observation. Missing data randomness can be divided into three sections:

- Missing completely at random: no difference between the characteristics of those missing and those that are not missing tested with little's MCAR [4] test e.g. lab or data entry mistakes.
- Missing at random: there is a difference but it can be explained using the existing data (not missing data) e.g. people with no income are younger than the ones with income.
- Missing not at random: the difference cannot be explained because it depends directly on the missing data e.g. probability of reporting income depends on level of income.

Missing data can cause the estimate to deviate systematically from the quantity of interest; this problem doesn't occur with missing completely at random data, but can occur in the other two types, another possible problem is the wrong standard error estimates, which leads to wrong study conclusions about the relationship between the predictive variables and the outcome variables. We should also mention that the impact of these problems depends on the quantity and the mechanism of these missing data.

2. The missing data treatments used in the comparison

The variety of missing data treatments requires the comparison, the statistician finds multiple choices to deal with the missing value problem, and therefore assessing their performances is a necessity to decide which approach is most efficient.

NIPALS algorithm

First presented by WOLD (1966), with the name NILES is an algorithm that performs principal component analysis on data sets that contain missing cases using an iterative procedure. the main idea is to calculate the slopes of the least squares line that crosses the origin of the points of the observed data; in this case the eigenvalues are determined by the variance of the NIPALS components; the same algorithm can estimate the missing data, but it can function without having to estimate them. The convergence of the algorithm depends on the percentage of missing data [6], in this paper we used the software R, specifically the package plsdepot developed by Gaston Sanchez [7], here is the detailed description of the algorithm as presented by Tenenhaus [6]:

Step1:$X_0 = X$

Step2: for $h = 1,2, \dots, a$

Step 2.1:$t_h =$ first column in X_{h-1}

Step 2.2: repeat until convergence ofp_h:

Step 2.2.1:$j = 1,2, \dots, p$

$$p_{hj} = \frac{\sum_{\{i:x_{ji} \text{ and } t_h \text{ non missing}\}} x_{h-1,ji} t_{hi}}{\sum_{\{i:x_{ji} \text{ and } t_h \text{ non missing}\}} t_{hi}^2}$$

Step 2.2.2: normalize p_h to 1

Step 2.2.3: $i = 1,2, \dots, n$:

$$p_{hj} = \frac{\sum_{\{i:x_{ji} \text{ and } t_h \text{ non missing}\}} x_{h-1,ji} t_{hi}}{\sum_{\{i:x_{ji} \text{ and } t_h \text{ non missing}\}} t_{hi}^2}$$

Step 2.3: $X_h = X_{h-1} - t_h p_h'$

Expectation maximization with bootstrap (EMB)

The expectation maximization algorithm was originally presented in 1977 by Dempster and al [5]; it's an algorithm that determines the maximum likelihood estimator of an unknown distribution parameter which can be used to impute missing cases with predictive distribution values. Suppose we have a density $f(.,\theta)$ with an unknown parameter, and a set of incomplete data $X = (X_{obs}, X_{miss})$ where X_{obs} and X_{miss} refer to observed values and missing values (respectively), in this algorithm we maximize the likelihood of the observed data by using the sample X, every iteration $l = 1, 2 \dots$ of the algorithm consists of two steps that are repeated until convergence [5]:

Expectation step: we calculate the conditional expectation$Q(\theta, \theta^l)$ of $L(\theta, X)$ relatively to the density $f(X_{miss}/X_{obs}, \theta^l)$.

Maximization step: we chose $\theta^{(l+1)}$ such as $Q(\theta^{(l+1)}, \theta^l) \geq Q(\theta, \theta^l)$for all θ.

In this paper we used an improved version of this algorithm developed by James Honaker, Gary King and Matthew Blackwell [8], in an R package called Amelia; the package uses the classical EM algorithm on different bootstrapped samples, then it combines the outcome in

the multiple imputation. It assumes that the data set has multi-normal distribution and is missing at random.

Mean imputation

It consists of filling the missing value with the mean of the non-missing values of the same variables; if used on missing completely at random data the mean imputation doesn't compromise the result of the analysis, but it does if used the other two types (MCAR, MNAR), it can lead to a vast underestimation of the standard errors, it can also change the relationship between variables, despite all of its drawbacks, mean imputation is an appealing approach given its simplicity.

3. Comparison of the three methods

As we have mentioned before, missing data is a common problem in all kinds of surveys, which makes the statistician choose whether he should simply omit the missing values, to only use the observed ones, or should he use a more robust approach. We surely advocate imputation methods, since they are very efficient to assure the performance of the statistical analysis, in this paper we chose principal component analysis as the statistical method on which we compare three missing data treatments by using simple square differences of the eigenvalues.

Step 1: the first step consists of simulating a multinormal distribution data set with 100 observations and six variables; we used the package MASS [9] for the simulation, and we performed a PCA (with the FactoMineR package [10]) on the simulated matrix, and these were the eigenvalues:

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	2.6530999	44.218331	44.21833
comp 2	1.5380157	25.633594	69.85193
comp 3	1.0275408	17.125680	86.97761
comp 4	0.3623788	6.039647	93.01725
comp 5	0.2352939	3.921565	96.93882
comp 6	0.1836709	3.061182	100.00000

Step 2: we generate missing data in every variable of the simulated data set using a uniform distribution; we varied the percentage of missing values from 5 % to 20 % to test the limits and stability of every method when performing a principal components analysis.

Step 3: the third step consists of imputing the data set using EBM algorithm and mean imputation, performing a PCA on the imputed data and running the NIPALS algorithm; on each PCA result we calculate the indicator:

$$I = \sum_{i=1}^{6} (\alpha_i - \overline{\alpha_i})^2$$

Where α_i denotes the i_{th} eigenvalue of the PCA on the imputed matrix (EBM and mean imputation) or NIPALS result and $\overline{\alpha_i}$ the eigenvalue of the PCA conducted on the complete data set. This indicator is inspired by the one presented in [11], the goal of the indicator is to assess the importance of the gap between the imputed result and the original one. We chose to compare eigenvalues due to their importance in the PCA, and because of their reduced dimensionality compared to the data set.

Results of the comparison

In order to view the outcome of the comparison properly we will present the plot of the variation of the indicator for all three methods:

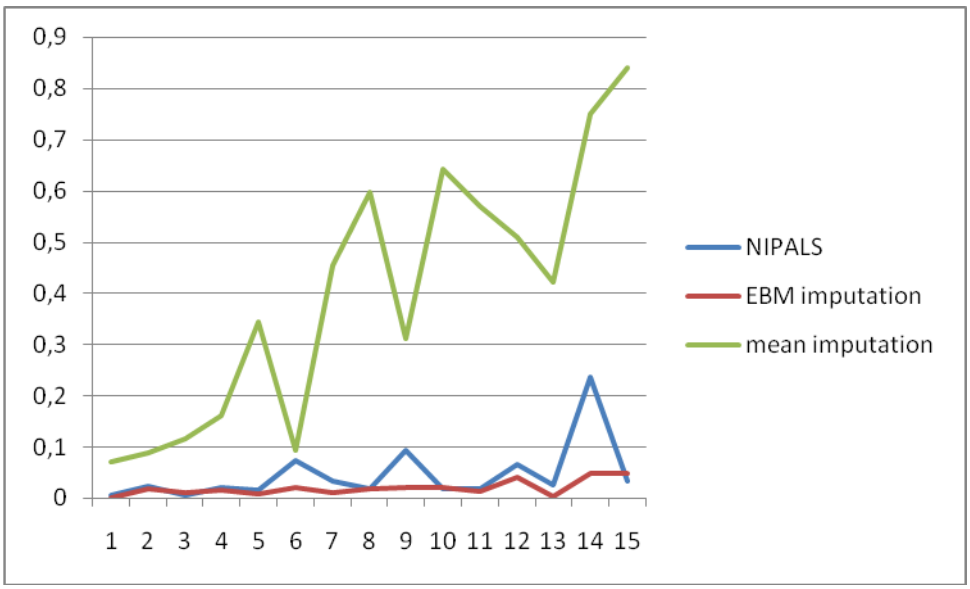

Figure 1: the evolution of the indicator

The indicator seems to be more stable when using the EBM algorithm, with a maximum value of 0.047 when we have 20% of missing value, which is almost negligible; on the other hand the NIPALS algorithm seems to be less stable but eventually has the same trend as the EBM

algorithm with a maximum value of 0.2 which is considered a vast gap from the real eigenvalues. Finally, the mean imputation seems to be the least effective as it continues to diverge with the increase of missing data.

4. Result and discussion

This paper presented three missing data treatments in a comparison approach, the criterion used in the comparison is an indicator based on the eigenvalues of the principal component analysis on the imputed data sets, to test the limits of each method's performance we varied the percentage of missing values from 5% to 20%, the EBM algorithm seemed to be the best, since its imputed data had very similar eigenvalues to the one conducted on the complete data set. In one of our references [11], a similar comparison was conducted; the conclusion was that NIPALS was better in terms of the indicator's performance, which is different from our result. It is important to state that the comparison didn't include a data mining procedure in their case; it only consisted of comparing the imputed data with the original ones. As a conclusion, the imputed data can have a different influence in the modeling process regardless of its similarity to the original data set.

Bibliography

[1]I.B. Aydilek, A. Arslan (2012), *A novel hybrid approach to estimating missing values in databases using k-nearest neighbors and neural networks,* International Journal of Innovative Computing, Information and Control 8 (7A) (2012) 4705–4717.

[2]D. Li, J. Deogun (2004), W. Spaulding, B. Shuart, *Towards missing data imputation: a study of fuzzy K-means clustering method,* Rough Sets Curr. Trends Comput. 3066 (2004) 73–579.

[3]F.V. Nelwamondo, S. Mohamed, T. Marwala (2007), *Missing data: a comparison of neural network and expectation maximization techniques,* Curr. Sci. India93 (2007) 1514–1521.

[4] Roderick J. A. Little (1988), *a test for missing completely at random,* American statistical association.

[5] A. P. Dempster, N. M. Laird and D. B. Rubin (1977), *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society.Series B (Methodological), Vol. 39, No. 1 (1977), pp. 1-38.

[6] Michel Tenenhaus (1998), *régression pls théorie et applications,* édition technip, Paris.

[7] Gaston Sanchez (2013), *package 'plsdepot',* R (>= 2.15.1).Version 0.1.17, 2012-11-12.

[8] James Honaker, Gary King and Matthew Blackwell (2011), *Amelia II: A Program for Missing Data*, journal of statistical software.December 2011, Volume 45, Issue 7.

[9] Brian Ripley, Bill Venables, Douglas M. Bates, Kurt Hornik, Albrecht Gebhardt, David Firth (2014), *package 'MASS' version: 7.3-31,*R (>= 3.0.0).Version 7.3-31, March 28, 2014.

[10] SébastienLê, JullieJosse, François Husson (008), *FactoMineR: an R package for multivariate analysis,* journal of statistical software.March 2008, Volume 25, Issue 1.

[11] CristianPreda, Alain Duhamel, Monique Picavet, Tahar Kechadi, *gestion des données manquantes dans les grandes bases de données en santé,* journées francophones. d'informatique Médicale, Lille 12-13 Mai 2005.

[12] J.Ross Quinlan (1988), *C4.5: programs for machine learning*, university of Sydney.